# ChromatoGate 1.2 Manual

Nikolaos Alachiotis
The Exelixis Lab
Heidelberg Institute for Theoretical Studies
n.alachiotis@gmail.com

Emmanuella Vogiatzi
Institute of Marine Biology and Genetics
Hellenic Centre for Marine Research
evogiatzi@her.hcmr.gr

## 1. What is ChromatoGate

ChromatoGate (CG) has been created to accelerate the process of detecting possible errors in DNA sequences that have been introduced by Sanger sequencers. To detect possible errors in the sequences, CG starts from the multiple-sequence alignment instead of inspecting every sequence separately. Therefore, the error-detection procedure is somehow reversed. For the straight-forward one, that is, the per-sequence error inspection/correction prior to alignment, the Phred tool can be used. CG does not align nor changes anything in the sequences, that is, it does not remove any possible sequencing errors. It implements a series of steps required in the multiple-sequence alignment generation process. During this CG-guided alignment generation procedure, the tool gathers information about alignment gaps, trimmed sequence edges, forward/reversed/consensus sequences, and corrections that have been already applied to the sequences by the user. Using all gathered information, CG is able to find for every base in the sequence alignment the corresponding chromatogram peak. For all the bases which the user considers to be responsible for degrading the quality of the alignment, CG points out which chromatogram peaks correspond to them and the user can easily assess whether they were called correctly by the sequencer's base call software. CG automatically calculates the chromatogram peak positions for all ambiguity characters in an alignment as well as for all bases responsible for categorizing a site as polymorphic based on a user-defined sensitivity criterion.

## 2. Source Code and Installation

The source code of ChromatoGate is sequential C since it does not execute any highly compute-intensive operations. The most computationally intensive part of the CG is the consensus sequence generation. The pair-wise alignments are carried out using the Smith-

Waterman algorithm and due to the short sequence lengths, the generation of a consensus sequence terminates in seconds. However, if there are many forward/reversed sequence pairs to generate consensus sequences for, CG might take some minutes for the consensus operation.

There are two code versions, one for Windows and one for Linux platforms.

WINDOWS: An executable (.exe) file is provided.

```
LINUX: Compile the source code with:     gcc ChromatoGate.c -o ChromatoGate -lm
       Start ChromatoGate with:          ./ChromatoGate
```

## 3. How to use ChromatoGate (step-by-step)

1. Extract the FASTA sequences from the chromatograms. An easy-to-use open tool than can do that (among several other things) is BioEdit. Open the chromatograms and extract the sequences in FASTA format without any modification, that is, do not correct any obvious sequencing errors. If you want to trim an edge you can do it now. In order to trim an edge in a CG-compliant way you must replace some (more than 2) bases with gaps. We call this short gap sequence TI (Trim-Indicator).

   Example:

   ```
           initial sequence:      NNNCGNNNACAGGTCGGGTC
           sequence with TI:      NNNCG - - -  ACAGGTCGGGTC
   ```

   CG will throw away the subsequence NNNCGNNN. Every FASTA sequence must have at most one TI at the beginning and at most one at the end.

2. Run ChromatoGate. You will be prompted to enter a name for the workspace. Enter a name and press enter. CG will generate the workspace and all the required folders in it. Then you will see the CG function list.

3. Before selecting a CG function from the list, put the FASTA files that contain sequences amplified with a forward primer in the FORWARD folder of the workspace and those amplified with a reverse primer in the REVERSE folder.

4. Go back to the function list. If you want to calculate consensus sequences select the Consensus Sequence Generation function (CSG). Otherwise proceed to the next step. If you select the CSG function, a folder CONSENSUS will be created in the workspace. The CONSENSUS folder contains two folders: FORWARD and REVERSE. Put the FASTA files that contain sequences amplified with a forward primer in the FORWARD folder and those amplified with a reverse primer in the REVERSE folder. Rename the sequences in the FORWARD or the REVERSE folder so that, those files in the two folders that contain sequences which should be combined by CG in order to create a consensus must have the same name. When the sequences are in the CONSENSUS/FORWARD and CONSENSUS/REVERSE folders press enter. Then you will be prompted to choose between two alignment scoring functions. These values are

used by the Smith-Waterman algorithm for the alignment. The default values have been chosen experimentally. The second option is to provide user-defined values.
The next prompt asks for a mismatch handling strategy. During the consensus sequence extraction, AMB strategy replaces mismatches with the respective ambiguity codes while N strategy replaces mismatches with N. The consensus sequences will be generated in the CONSENSUS folder in FASTA format. Press enter to terminate CG.

5. Run ChromatoGate again and select the same workspace. Execute the Preliminary File Generation Option (PFG). The FASTA files should already be in the FORWARD, REVERSE, and CONSENSUS folders. Press enter to generate the preliminary file and then enter again to terminate CG.

6. Use the preliminary file as input to any multiple-sequence alignment program and generate the alignment. Save the alignment in the workspace.

7. Run ChromatoGate again and select the same workspace. Execute the Ambiguous Character Detection (ACD) function and enter the name of the alignment file. This will generate one (or more) reports in the workspace. Every entry in an ACD report corresponds to an ambiguity code in the alignment and consists of fours fields: Sequence name, Alignment site, Chromatogram position, Changed to. Sequence name and alignment site fields tell you where in the alignment the ambiguity code was detected. You can go to the respective chromatogram file, at the position indicated by the “Chromatogram position” field and assess whether you want to leave the ambiguity code as it is or replace it with a base. If you decide that the ambiguous character can be replaced with a base, you should apply that change to the sequence and change the question mark in the “Changed to” field to the newly inserted base.

8. Once you are finished with the correction of the ambiguity codes and the update of the ACD reports accordingly you must degap the multiple-sequence alignment and realign. This can also be done with BioEdit. Save the alignment in the workspace.

9. After the second alignment, run ChromatoGate once more and select the same workspace. Execute the Polymorphic Site Detection (PSD) function and enter the name of the new alignment file. Then you will be prompted to enter a sensitivity value. You can either enter an integer value or a percentage. This value is used by CG to determine the polymorphic sites. After that you will be prompted to select a search strategy: Less or Equal (LE) or One Major Class (1MC). The LE strategy selects sites which contain characters that appear less than or equal times to the sensitivity value. The 1MC strategy selects sites which contain characters that appear less than or equal times to the sensitivity value and all remaining characters in the site belong to the same nucleotide class. The PSD function will generate a PSD report containing an entry for every polymorphic site. Each entry consists of four fields: Alignment site, Site report, Sequence name and Chromatogram position. The site report shows the statistics for the current site. For every character responsible for polymorphism in the site, the “Sequence name” tells you to which sequence the character belongs and the Chromatogram position shows which peak or peaks correspond to the character. Every “Chromatogram position” entry consists of the primer, the position of the chromatogram peak, and the nucleotide found at the specific position.

10. Once you have gone through the PSD report, assessed the corresponding chromatogram peaks, and corrected possible errors the CG-based alignment generation and correction procedure is over.

## 4. File Format

All input sequence files must be in FASTA format. In addition, the sequence files must have the extension “fas”.

## 5. Edge Trimming

Trim-indicators” are used for the trimming of the input sequences during the initial gathering of the input sequences into the FASTA file. A “trim-indicator” is a sequence of gaps (typically two or more). The sequence of gaps can be manually inserted into the input sequence by replacing an equal number of nucleotides that belong to the edge to be trimmed.
Note that, total absence of “trim-indicators” still yields a valid input sequence thus allowing for using the sequence exactly as it is when it comes out of the sequencer.

| | |
|---|---|
| Example of the initial sequence: | >seq name ACGTCGCTCGATATCGATCGCTAG |
| Sequence with “trim-indicators”: | >seq name ACG - - - CTCGATATCGATCG- - AG |
| Sequence with the edges trimmed: | >seq name CTCGATATCGATCG |

## 6. Consensus Sequence Generation – CSG

The CSG function is used to generate consensus sequences. The user must provide as input two sequences, one amplified by a forward primer and one amplified by a reversed primer. In the background, information regarding the origin of every nucleotide in each consensus sequence is stored in order to be used by the ACD and PSD functions.

1.1 PROMPT
Initially, the tool prompts the user to place the input sequence files into the CONSENSUS \ FORWARD and CONSENSUS \ REVERSE folder according to the primer each sequence has been amplified with. Input sequences must comply with the supported sequence and filename formats: sequences in FASTA format, filenames with extension “.fas”.

Trim-indicators (few-gap sequences) can appear in the input files. Sequences in the CONSENSUS \ REVERSE folder -MUST NOT- represent the reversed complement of the original, sequencer-generated, sequences. Apart from the sequence and filename requirements, an additional CSG filename requirement exists: For every file in the CSG FORWARD folder (e.g., example seq.fas) the file in the REVERSE folder that will be used in the generation of the consensus sequence must have exactly the same name: example seq.fas.

1.2 PROMPT

The Smith-Waterman algorithm is used for the alignment of the two input sequences for every consensus sequence generation. The second prompt allows the user to select between the default alignment scoring function or insert manually user-defined integers for the match score and mismatch and gap penalties. The respective default values have been decided through experiments on real-world sequences and are as follows: MATCH=3, MISMATCH=-3, GAP=-3

1.3 PROMPT
Through the third prompt the user can select a mismatch handling strategy (MHS). The MHS decides which ambiguous character will appear in the final consensus sequence at the positions that the alignment algorithm has detected a mismatch. The AMB MHS places the closest ambiguous character while the N MHS uses the N character for all mismatches.

## 7. Preliminary File Generation – PFG

The PFG function can be used to generate a preliminary FASTA file. The sequences in this file are not aligned. The user can calculate a multiple alignment of these sequences using any alignment program. In the background, information regarding the origin of every base in each input sequence is stored in order to be used by the ACD and PSD functions.

PROMPT
The tool prompts the user to place the input sequence files into the FORWARD and REVERSE folders of the selected workspace according to the primer each sequence has been amplified with. Input sequences must comply with the supported sequence and filename formats: sequences in FASTA format, filenames with extension “.fas”. Trim-indicators (few-gap sequences) can appear in the input files. Sequences in the REVERSE folder -MUST NOT- represent the reversed complement of the original, sequencer-generated, sequences. Consensus sequences that have been generated using the CSG function and thus already exist in the CONSENSUS folder are included in the preliminary file automatically.

## 8. Ambiguous Character Detection - ACD

The ACD function can be used to target the ambiguous characters in an input multiple alignment file that has been calculated from the preliminary file generated by the PFG function. Using information gathered during the CSG and PFG processess, the ACD function creates reports that contain the positions of each ambiguous character in the input alignment as well as the respective chromatogram positions in order to allow the user to correct possible errors.

PROMPT
The tool prompts the user to place the input alignment file into the workspace directory. Input sequences must comply with the supported sequence and filename formats: sequences in FASTA format, filename with extension “.fas”. Optionally, the input alignment file can contain sequences that were not present in the PFG-generated preliminary file. In this case, it is assumed that no chromatograms are available and thus chromatogram positions for nucleotides that belong to these sequences are not calculated. If there are chromatograms for these sequences repeat the PFG including them.

The ACD function generates an ACD report for every ambiguity in the alignment. Each report consists of four columns of data denoted as:

1. Sequence name
2. Alignment site
3. Chromatogram position
4. Changed to

The chromatogram position format is X.Y with X refering to the type of primer the respective sequence has been amplified with while Y being the chromatogram position. For sequences retrieved from the FORWARD folder X = FW. For sequences retrieved from the REVERSE folder X = RV. For sequences retrieved from the CONSENSUS folder the chromatogram position appears as FW.Y - RV.Y. The "Changed to" field in all reports is formed as: '>?'. The aim of existence of this field is described in the PSD section.

## 9. Polymorphic Site Detection – PSD

The PSD function is used to detect all the polymorphic sites in an input alignment based on a user-defined criterion. A PSD report that contains chromatogram positions of the nucleotides responsible for polymorphic sites in the alignment is generated in order to allow the user to correct possible errors introduced in the sequence by the sequencer. Furthermore, a post-PSD ACD report comprising all the ambiguous characters in the input alignment is generated.

4.1 PROMPT
The tool prompts the user to place the input alignment file into the workspace directory. This alignment can either be the same one used as input in the ACD function or a newly calculated one. If the ACD input alignment has been corrected based on the ACD reports, it is advisable to degap and realign the sequences and use the new alignment as input to the PSD function.

Input sequences must comply with the supported sequence and filename formats: sequences in FASTA format, filename with extension ".fas". Optionally, the input alignment file can contain sequences that were not present in the PFG-generated preliminary file or in the input alignment file to the ACD function. In this case, it is assumed that no chromatograms are available and thus chromatogram positions for bases that belong to these sequences will not be calculated. If chromatogram files for these sequences exist, repeat PFG and ACD functions including them.

4.2 PROMPT
The second prompt allows the user to enter a sensitivity value. This number is used in the selection of the polymorphic sites. The input can either be an absolute value, meaning that it must be less or equal to the number of sequences in the input alignment, or, a percentage. If the input value is a percentage it must be followed by '%' otherwise it is treated as an absolute value.

4.3 PROMPT
This prompt allows the user to select search strategy. There are two options: 1. Less or Equal (LE) and 2. One Major Class (1MC).The search strategy determines the way the sensitivity value is used to detect possible polymorphic sites.

1. LE: Selects the sites which contain characters that appear less or equal times than the sensitivity value.

- Example: A column with 50 As, 50 Cs, 50 Gs and 2 T is considered polymorphic if the sensitivity is equal or greater than 2 (absolute value).

2. 1MC: Selects the sites which contain characters that appear less or equal times than the sensitivity value and all remaining characters belong to the same nucleotide class.

- Example: A column with 150 As, 1 C and 1 G is considered polymorphic if the sensitivity is equal to 2.

The main difference between the alternative search strategies is that the 1MC strategy searches for one major nucleotide class in a site before applying the sensitivity-based criterion while the LE allows the existence of more than one major classes. If errors have been corrected in the sequences by the replacement of ambiguous characters with nucleotides then the user must update the respective entries in the ACD reports by replacing the symbol ’ ?’ in the “Changed to” field with the correction. Every replacement of the ’ ?’ symbol in any ACD report must correspond to a replacement of an ambiguous character in the alignment with another base. If one does not want to correct an ambiguous character in the alignment, the respective ACD report entry must not be updated. In that case, the respective ’ ?’ symbol in the ACD report will be treated as the pre-existing ambiguous character at the corresponding position in the sequence.

- REMEMBER - When reversed-primer amplified sequences are concerned the ’ ?’ symbol needs to be replaced with the complementary base of the one that appears in the chro matogram.

The PSD function generates a PSD report that contains an entry for every polymorphic site in the alignment. The report consists of four columns of data denoted as:

1. Alignment site
2. Site report
3. Sequence name
4. Chromatogram position

The site report provides the number of nucleotides the site consists of. The chromatogram position format is X.Y.Z with X again refering to the type of primer the respective sequence has been amplified with, Y being the chromatogram position and Z being the nucleotide at that position. For forward-primer amplified sequences Z represents the nucleotide in the sequence as well as the chromatogram while for reversed-primer amplified sequences Z is the complement of the nucleotide in the sequence in order to match with the nucleotide in the chromatogram. Furthermore, a post-PSD ACD report containing a list of all the remaining ambiguous characters in the alignment is generated.